Predictive Modeling using SAS Enterprise Miner™ and SAS/STAT®:

Principles and Best Practices

CAROLYN OLSEN & DANIEL FUHRMANN

1

Overview

This presentation **will**:

- Provide a brief introduction of how to set up analytics projects in SAS Enterprise Miner
- Highlight some of the main statistical concepts applied in a predictive modeling project
- Outline three supervised learning techniques

It **will not**:

- Offer in-depth coverage of advanced statistics
- Be a complete tutorial for analytics in SAS EG or SAS EM
- Be comprehensive by any stretch of imagination. Consider this presentation a starting point for your analytics journey.

Outline

1. Introduction
2. Business questions & analysis process
3. SEMMA methodology & project work flow
4. Considerations for predictive modeling
5. Selected supervised learning techniques
6. Model Assessment

introduction

1

It all starts with...

1

The answers might come via...

Predictive modeling: The use of known, historical data and mathematical techniques (statistical algorithms) to develop models that predict future events.

Typically, the end result is to streamline decision-making and to create new insights that lead to better actions or outcomes. Therefore, predictive modeling is widely used in industry to more accurately answer business questions that improve performance, increase revenues, or reduce costs.

Predictive Modeling Use Cases in Insurance

1

- Response model
- Default model (claims model)
- Retention model
- Cross-sell / Up-sell model
- Fraud model
- Lifetime Value model
- Process Failure model
- ... and the list goes on.

business questions & analysis process

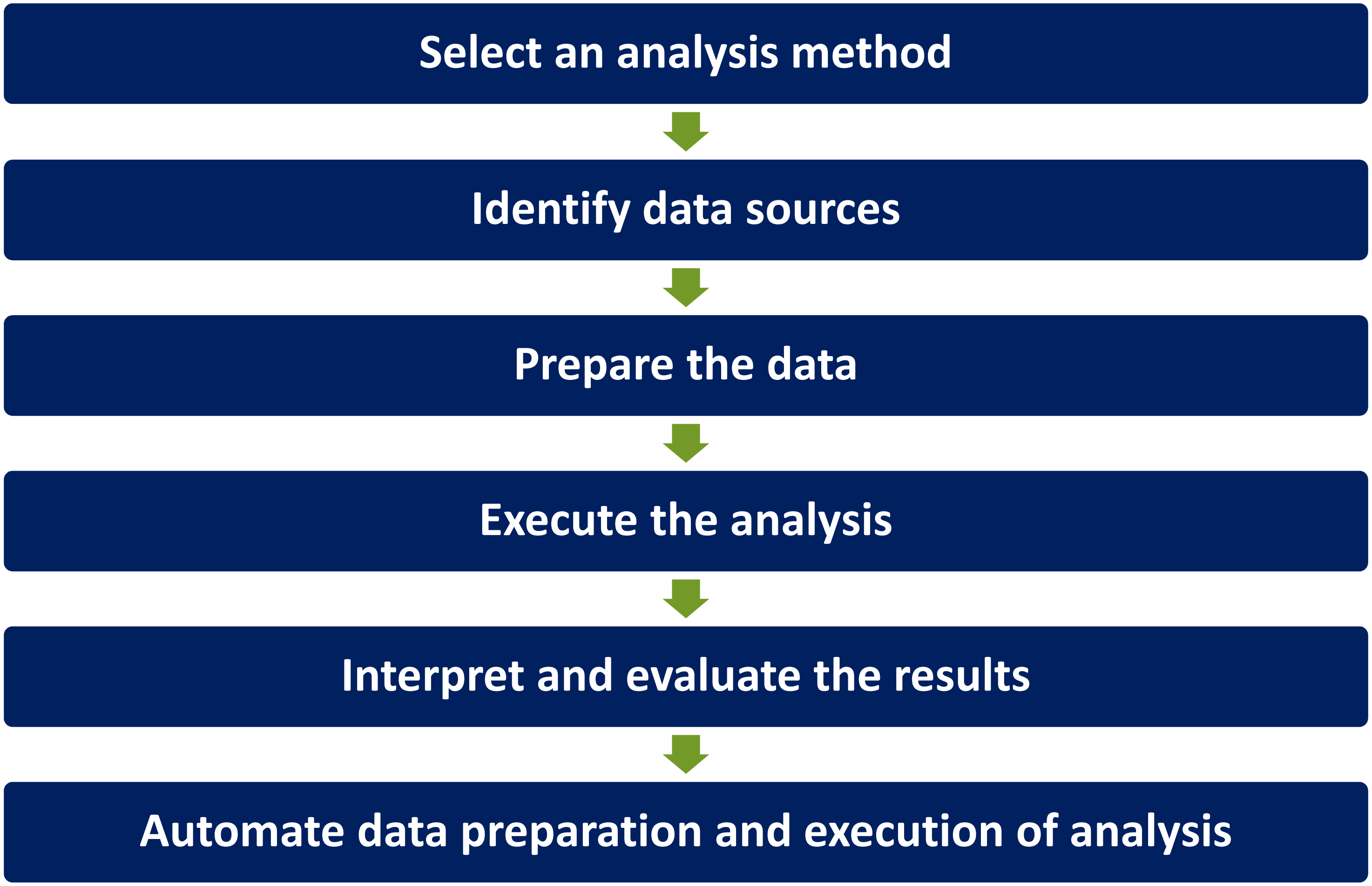
Characteristics of Analytic Business Questions

2

Analysis Complexity	• Descriptive or advanced analytics, or both?
Analysis paradigm	• Statistics or data mining?
Data preparation paradigm	• As much data as possible or business knowledge first?
Analysis method	• Supervised or unsupervised analysis?
Scoring needed	• Yes or No?
Periodicity of analysis	• One-time deal or re-run(s) anticipated?
Historic data needed	• Yes or No?
Data structure	• One row or multiple rows per subject?
Analytics team complexity	• Communication and documentation needed?

The Analysis Process (short)

2



Select an analysis method

Identify data sources

Prepare the data

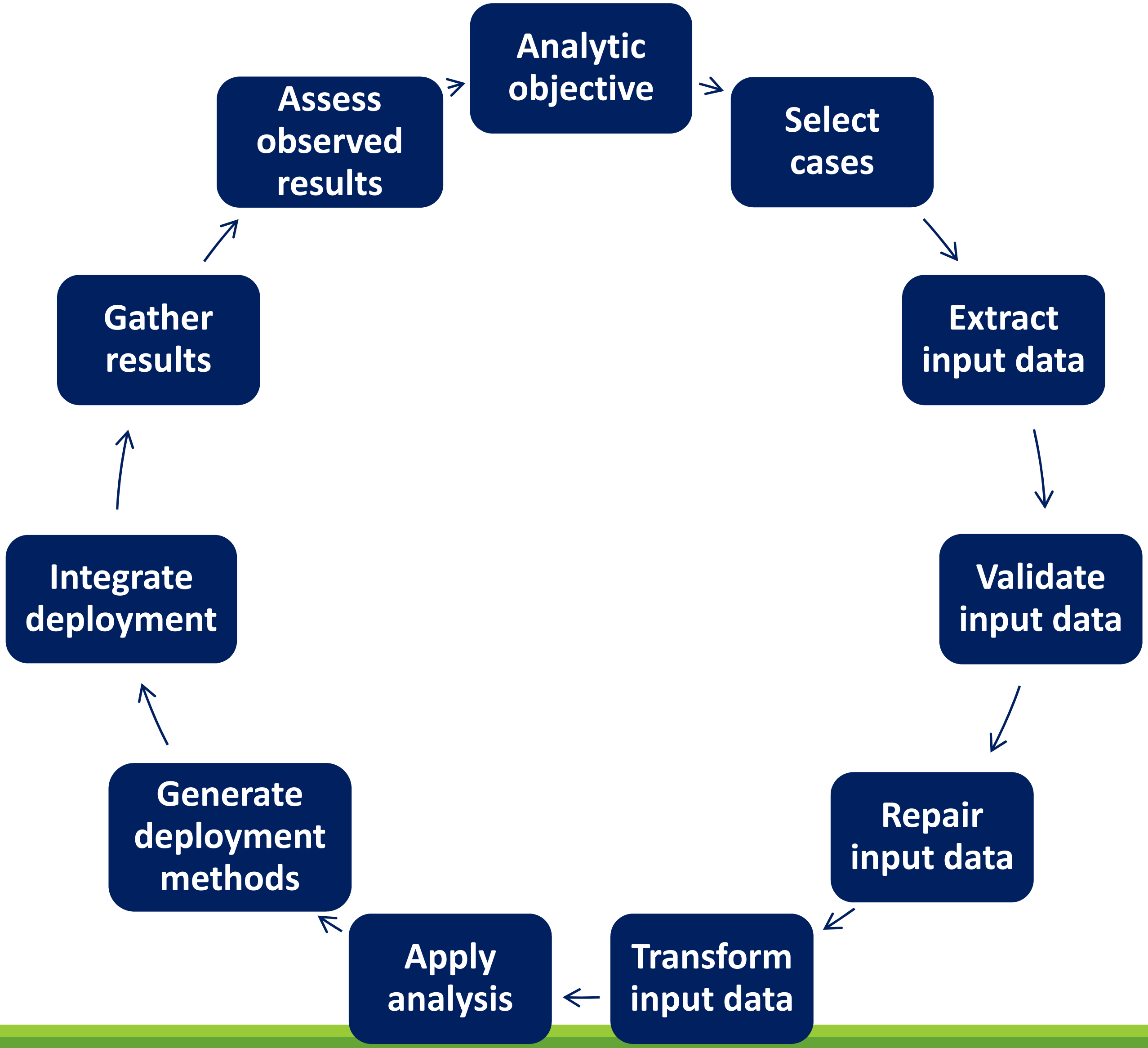
Execute the analysis

Interpret and evaluate the results

Automate data preparation and execution of analysis

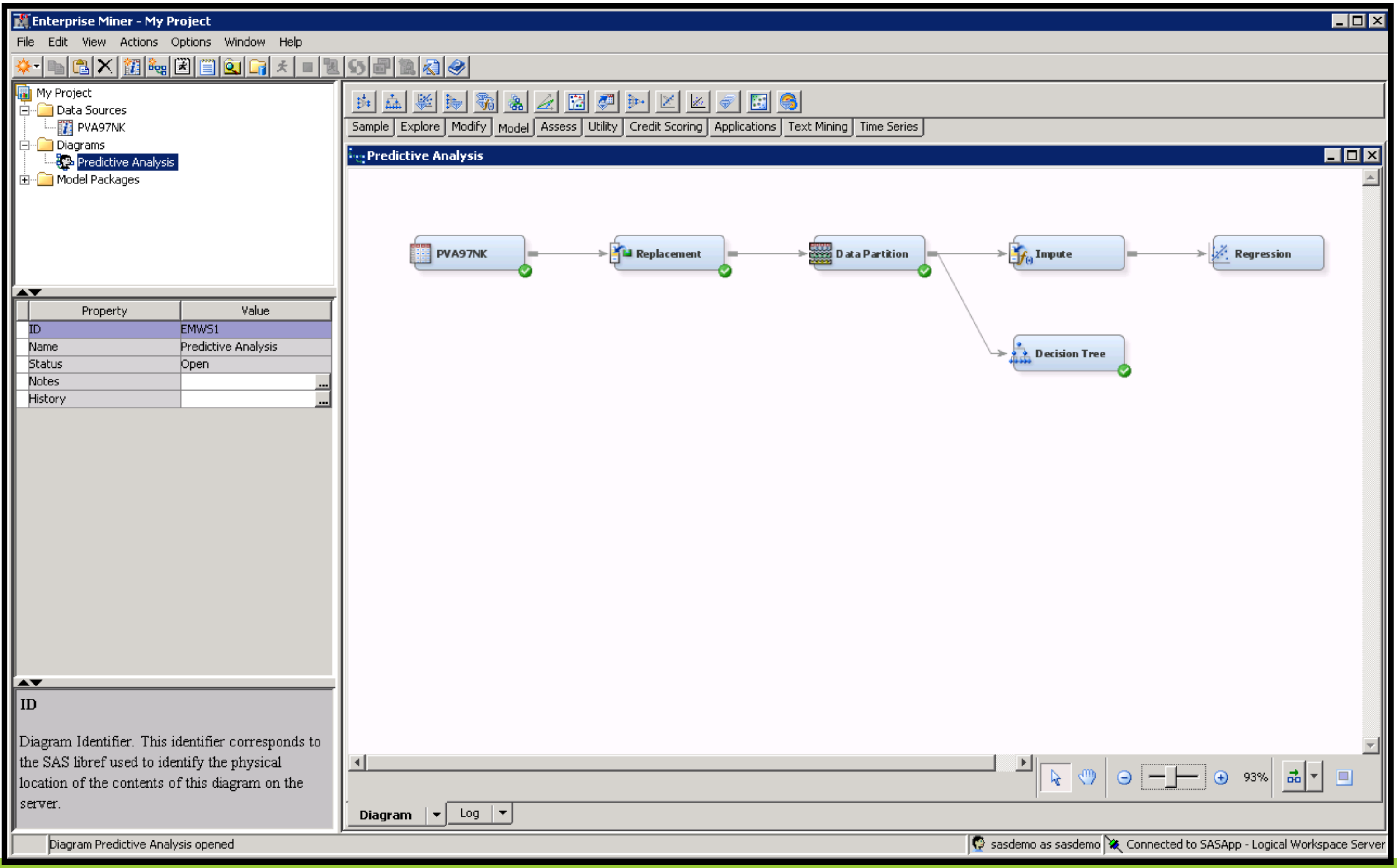
The Analysis Process (long)

2



semma methodology & process flow in SAS EM

Getting Started in SAS EM


Enterprise Miner - My Project
File Edit View Actions Options Window Help
My Project
Data Sources
PVA97NK
Diagrams
Predictive Analysis
Model Packages
Sample Explore Modify Model Assess Utility Credit Scoring Applications Text Mining Time Series
Predictive Analysis
PVA97NK
Replacement
Data Partition
Impute
Regression
Decision Tree
93%
Diagram
Log

Property	Value
ID	EMWS1
Name	Predictive Analysis
Status	Open
Notes	
History	

ID

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

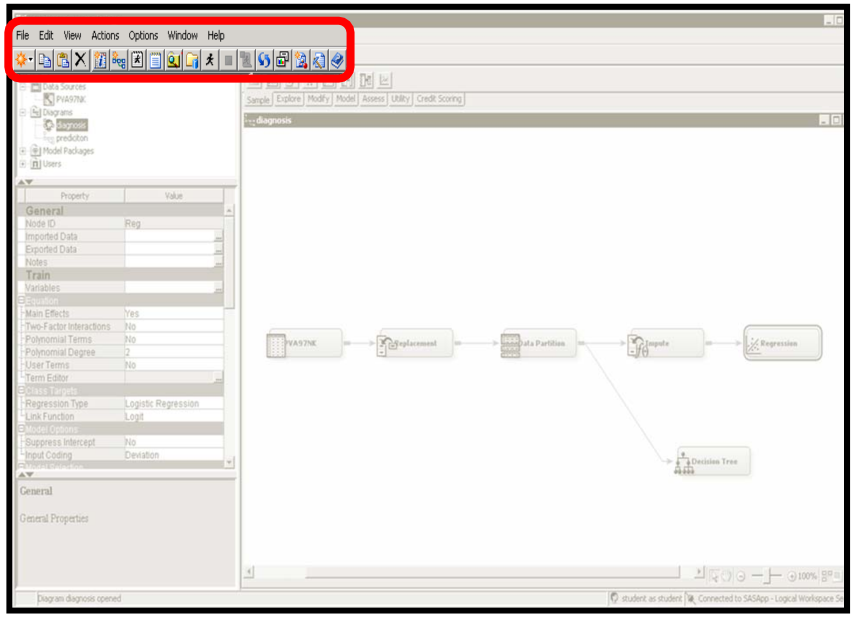
Diagram Predictive Analysis opened

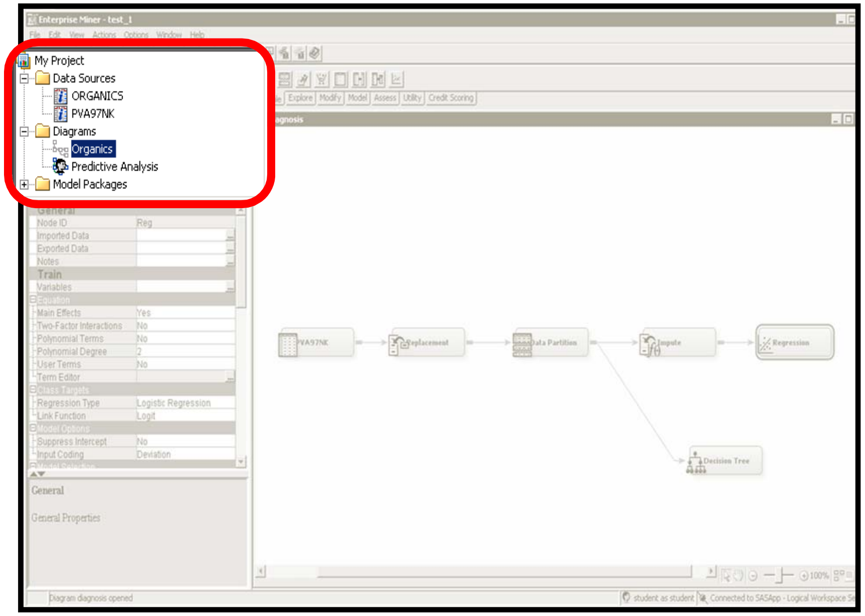
sasdemo as sasdemo

Connected to SASApp - Logical Workspace Server

3

Menu and Project Panel



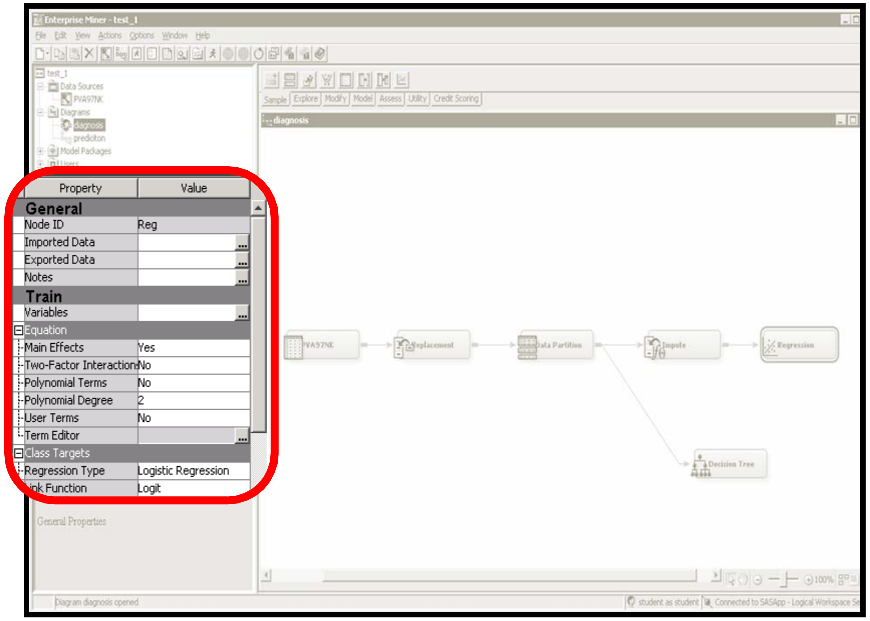


Menu/Task Bar and **Shortcuts**

Project Panel – Displays your data sources, diagrams, and model packages.

3

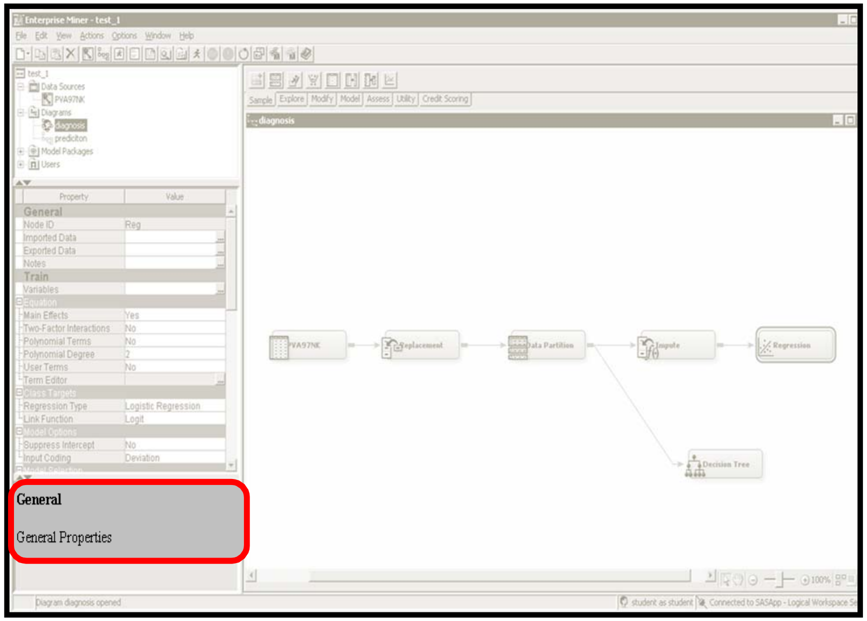
Properties and Help


Enterprise Miner - test_1
File Edit View Actions Options Window Help
test_1
Data Sources
PVA97NK
Diagrams
diagnosis
prediciton
Model Packages
Users
Sample Explore Modify Model Assess Utility Credit Scoring
diagnosis
PVA97NK
Replacement
Data Partition
Impute
Regression
Decision Tree
100%
student as student
Connected to SASApp - Logical Workspace Se

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟ Equation	
Main Effects	Yes
Two-Factor Interaction	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
⊟ Class Targets	
Regression Type	Logistic Regression
ink Function	Logit

General Properties

Diagram diagnosis opened


Enterprise Miner - test_1
File Edit View Actions Options Window Help
test_1
Data Sources
PVA97NK
Diagrams
diagnosis
prediciton
Model Packages
Users
Sample Explore Modify Model Assess Utility Credit Scoring
diagnosis
PVA97NK
Replacement
Data Partition
Impute
Regression
Decision Tree
100%
student as student
Connected to SASApp - Logical Workspace Se

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟ Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
⊟ Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
⊟ Model Options	
Suppress Intercept	No
Input Coding	Deviation
⊟ Model Selection	

General

General Properties

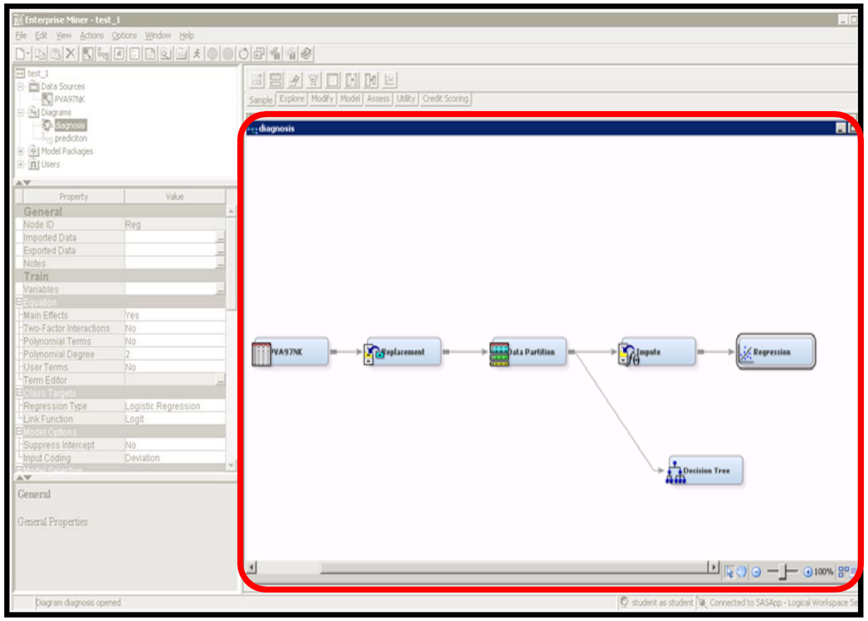
Diagram diagnosis opened

Properties Panel – User can make adjustments to Node properties here.

Help Panel – Displays brief yet helpful explanations of node properties and settings.

Diagram Workspace & SEMMA + Tools Tabs

3



Property	Value
General	
Node ID	Reg
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation

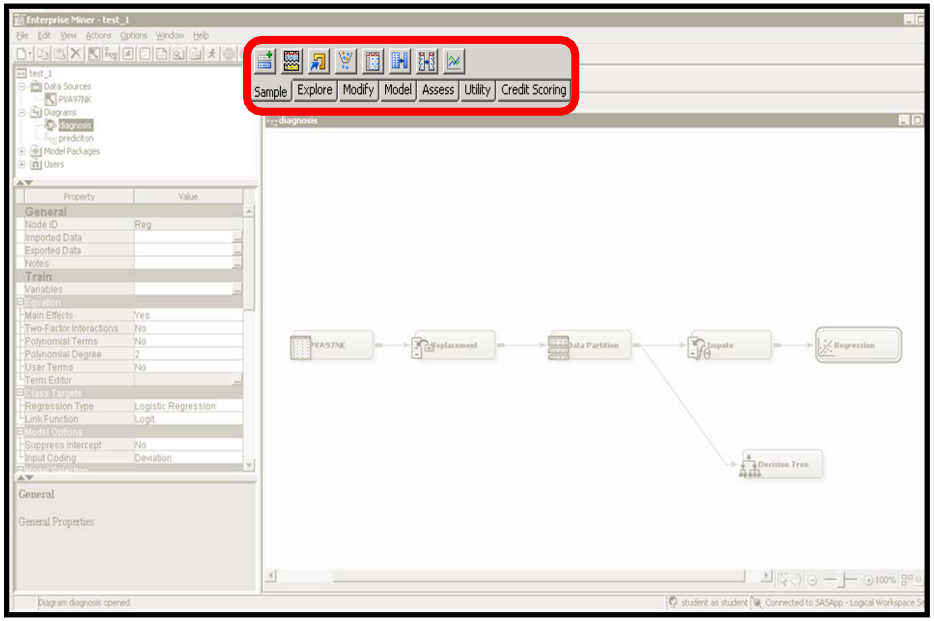
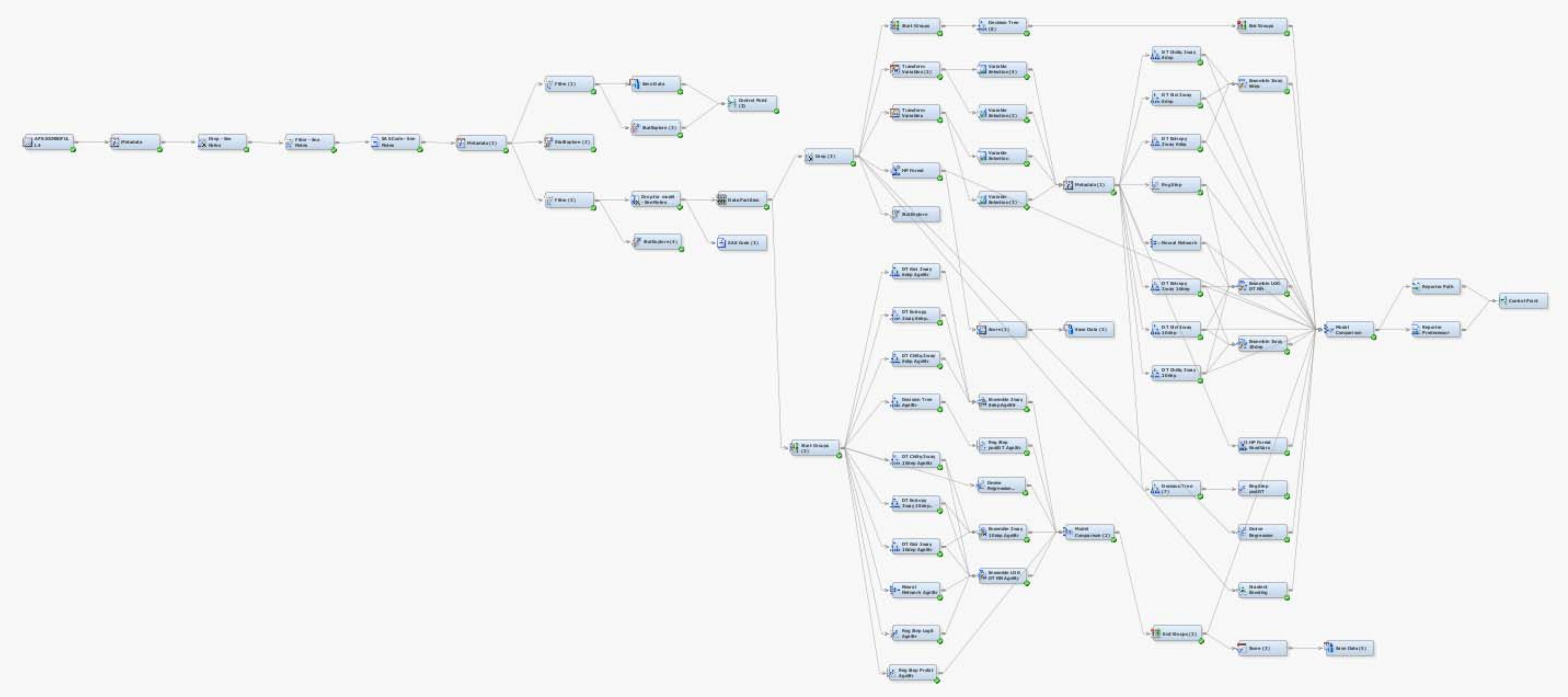


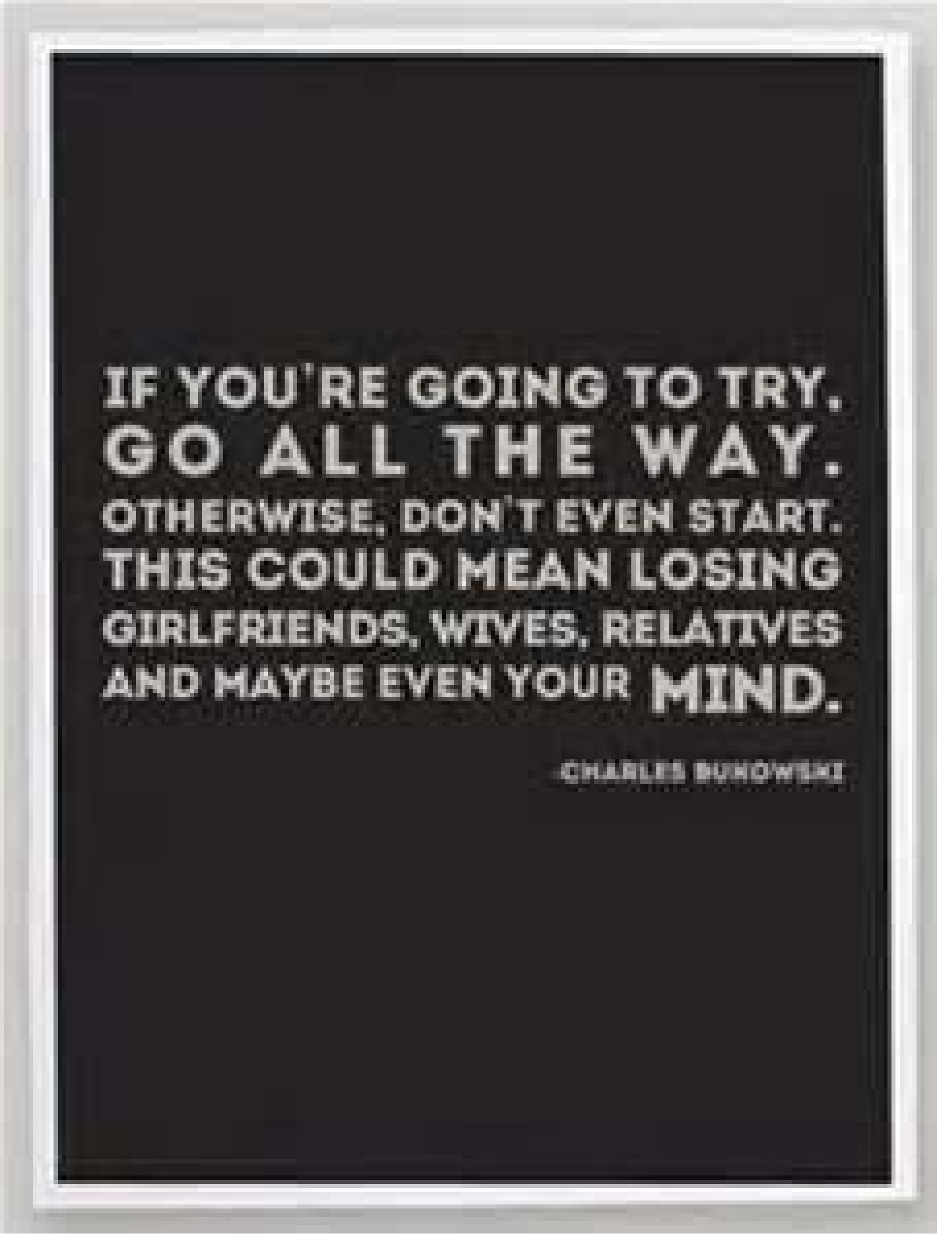
Diagram Workspace – User creates a logical process flow for the data mining project here. Information flows from node to node in the direction of the connecting arrows.

SEMMA & Tools Tabs– Functionalities to **s**ample, **e**xplore, **m**odify, **m**odel, and **a**sses are found here along with any additional licensed utilities (e.g. Text Miner, Credit Scoring, etc.)

3

Reality bites...





IF YOU'RE GOING TO TRY, GO ALL THE WAY. OTHERWISE, DON'T EVEN START. THIS COULD MEAN LOSING GIRLFRIENDS, WIVES, RELATIVES AND MAYBE EVEN YOUR MIND.

-CHARLES BUKOWSKI

3

considerations for predictive modeling

4

Outline

I) Introduction

2) Business questions & analysis process

3) **SEMMA methodology & project work flow in SAS Enterprise Miner**

4) Considerations for predictive modeling are outlined next, they include **data preparation (definitions of target and observation window, offset window, sampling, missing value treatment, and variable transformations), variable selection methods, overfitting, validation methods, and criteria to assess model performance.** This paper concludes with an **introduction to several supervised learning techniques (regression, decision trees, neural networks, and ensemble models) and applications of these methods to a use case in the insurance industry.** The intended audience for this paper are beginning and intermediate-level data scientists, analysts, and modelers. While the paper focuses on applications in SAS Enterprise Miner, it provides alternative approaches implemented in SAS/STAT, wherever possible.

Data Structures and Data Modeling – Things to consider

4

- The Origin of Data
- Data Models
- Analysis Subjects and Multiple Observations
- Data Mart Structures
- No analysis subject available?

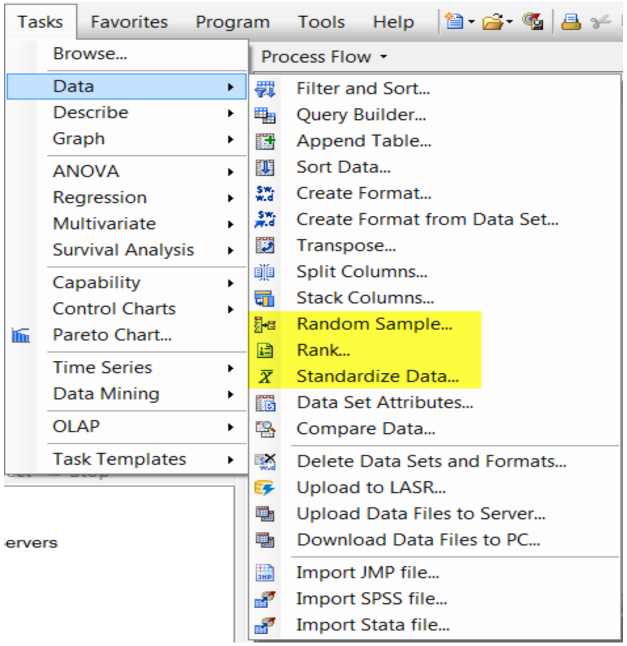
- For purposes of predictive modeling – it is important to decide the grain or event you want to predict.
  - i.e. the structure and design of your dependent variable(s)

Sample – How to

4

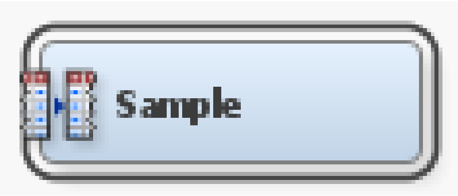
SAS ENTERPRISE GUIDE

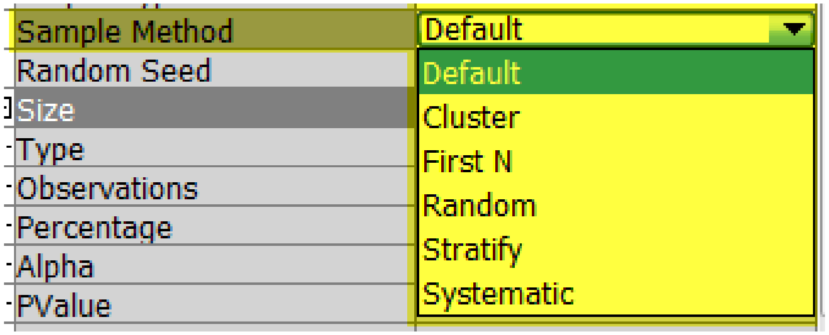
- PROC SURVEYSELECT
- PROC IML



SAS ENTERPRISE MINER

- SAS CODE Node
- Sample Node





Sample Method: Default

4

Definition of EDA

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set:

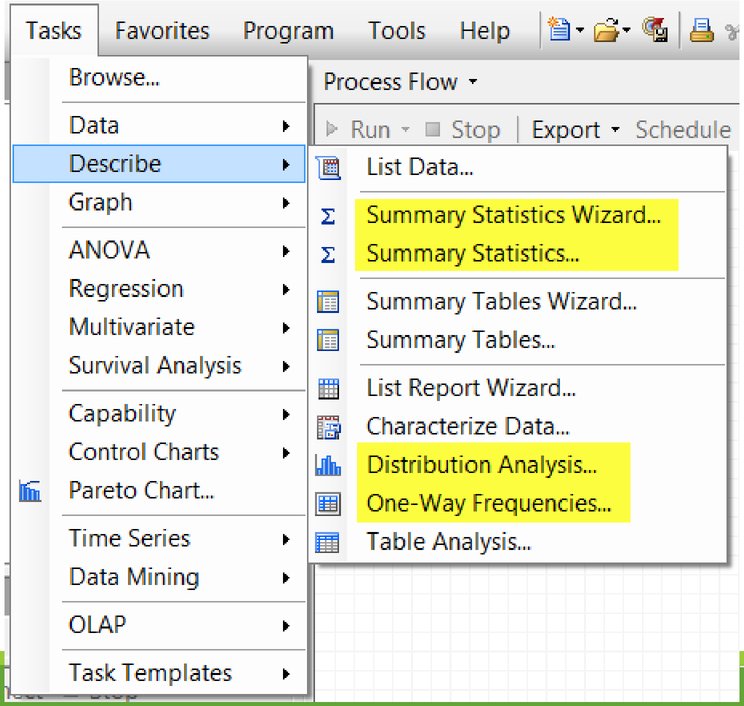
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- determine optimal factor settings.

Explore – How to

4

SAS ENTERPRISE GUIDE

- PROC UNIVARIATE & CORR
- PROC SGPLOT & MI



SAS ENTERPRISE MINER

- SAS CODE Node
- StatExplore, GraphExplore, Multiplot

4

Suggested Steps after EDA

Check for missing values (MCAR, MAR, MNAR)
→ Possible actions: imputation, leave as is, missing value flags, etc.

Look for univariate (and multivariate) outliers
→ Possible actions: drop obs, Winsorizing, value correction if erroneous entry, etc.

Investigate variable distributions for ill-behaved items
→ Possible actions: transformation, binning, etc.

Check if multicollinearity is present
→ Possible actions: drop input(s), create factors or principal components, do nothing

Discuss any existing business rules around data with IT Expert and Business Client, ideally before taking any of the actions above.

Modify - Set up of Independent Variables

4

Basic Variables

- Continuous or Interval Variables
- Categorical (nominal or ordinal)
  - Dummy Variables

Intermediate Variables

- Interaction Effects
- Transformed (Power, Logarithmic, etc.)
- Binning
  - Interval to Categorical Variables

Modify - Important Variable Criteria

4

Sufficiency

- All potentially relevant information from the available source systems is also in the analysis table.

Efficiency

- Keep number of variables as small as possible (parsimony).

Relevance

- Data are gathered/aggregated in such a form that the (derived) variables are suitable for the analysis and business question.

Interpretability

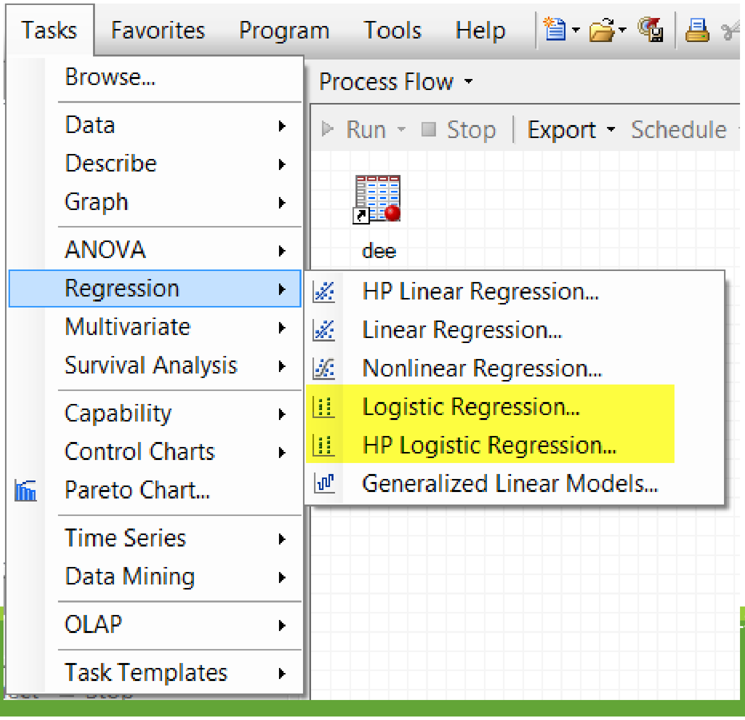
- Variables that are used for analysis can be interpreted and are meaningful from a business point of view.

4

Modify – How to

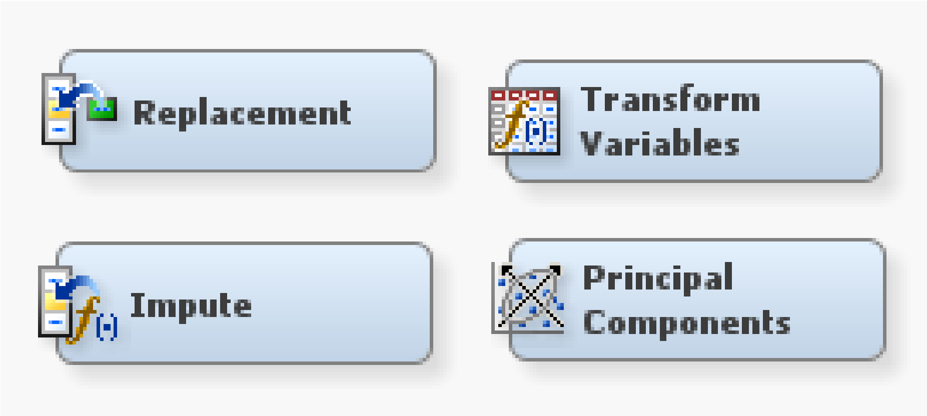
SAS ENTERPRISE GUIDE

- DATA step
- Specify interactions, polynomials within analytical tasks



SAS ENTERPRISE MINER

- SAS CODE Node



selected supervised learning techniques

5

Logistic Regression

PURPOSE

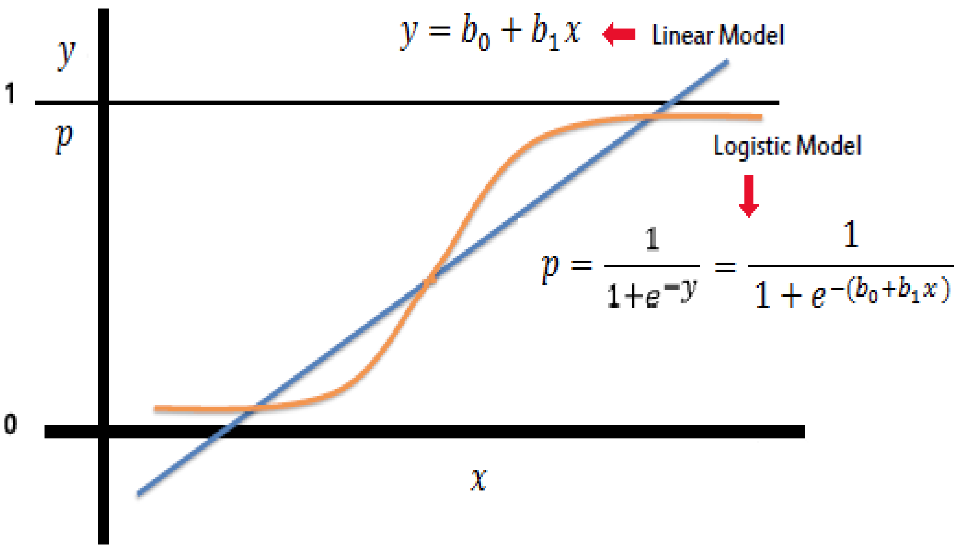
Use multiple *x* variables (inputs) to estimate *p* (output), the probability of an event *y*. This is a **supervised** technique that requires historical data with known outcomes.

For a probability *p* of an event, the odds of the event are *p*/(1-*p*). A logistic regression models the log of the odds of the event, using a linear function:

logit = log(p/(1-p)) = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_P X_P$

The coefficients in this equation are determined using “maximum likelihood estimation,” an iterative process that begins with a tentative solution then repeats revision until improvement is minute – at which point the process is said to have “converged.”

METHOD VISUALIZATION



5

Logistic Regression

IMPORTANT CONSIDERATIONS

There are assumptions which must be met for a logistic regression to be unbiased, and there are important considerations around data set structure and defining the outcome. Additionally, with small samples, maximum likelihood estimation is known to be biased.

Logistic regression has the same limitations in modeling non-linear relationships as linear regression. That is because the log odds is still estimated using a *linear* equation. This requires extra attention to transformations and interactions.

MODEL VALIDATION

There is a variety of measures used to measure how strong a classification model is: the **confusion matrix**, which gives true and false positive rates and true and false negative rates; **accuracy**; **misclassification rate**; **positive predictive value (PPV), negative predictive value (NPV)**; and others. Review the Receiver Operative Characteristic (**ROC**) curve.

For a logistic regression, you can also review a **Likelihood-Ratio Test** as a measure of goodness of fit; or you can review **AIC**, **BIC**, or one of several “**pseudo R^2**” values.

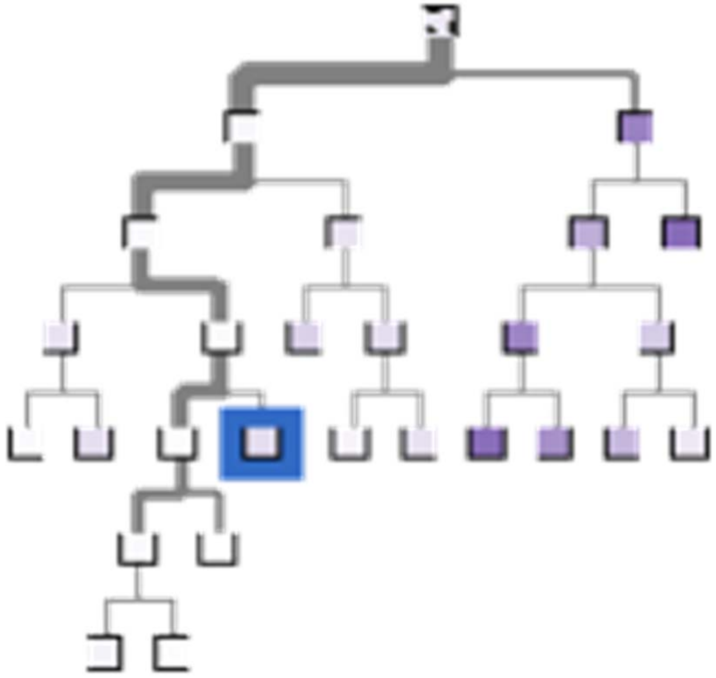
5

Decision Tree

PURPOSE

Use multiple *x* variables to predict a categorical outcome *y*. The outcome can be two-level (yes/no), or multi-level. These can be used either for predictive purposes, or for data exploration. This is a **supervised** technique that requires historical data with known outcomes.

METHOD VISUALIZATION



5

Decision Tree

IMPORTANT CONSIDERATIONS

Decision trees do not have as strong assumptions as regressions do, but they are many different ways to approach them: algorithm choice, missing value treatment, stopping criteria, pruning methods, etc. Additionally, the predictive power of decision trees is greatly increased by use of ensembling techniques.

SAS Enterprise Miner uses its own algorithm which adopts aspects of the various decision tree algorithms and allows the user to set parameters for missing variable handling and stopping criteria. Outside of SAS EM, you will need to choose an algorithm and handle your data preparation accordingly.

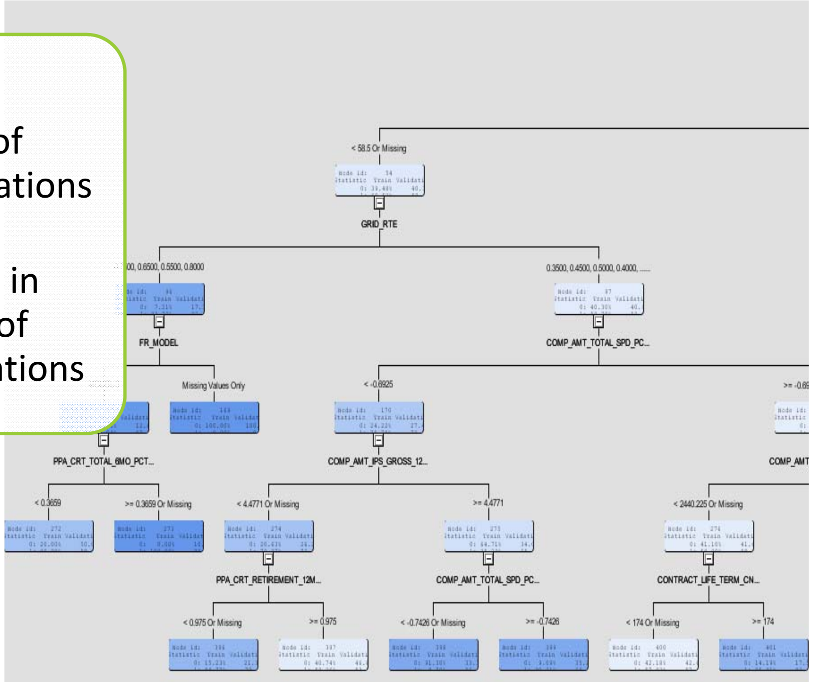
MODEL VALIDATION

There is a variety of measures used to measure how strong a classification model is: the **confusion matrix**, which gives **true and false positive rates and true and false negative rates**; **accuracy; misclassification rate; positive predictive value (PPV) and negative predictive value (NPV)**; and others. Often analysts will review the Receiver Operative Characteristic (**ROC**) curve to see how well the model is doing and to review possible trade-offs of sensitivity and specificity.

Decision tree output in SAS EM

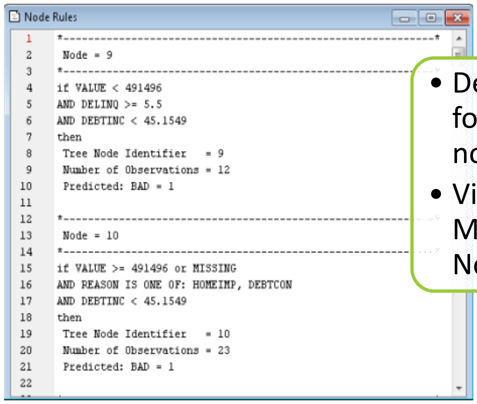
5

Tree Plot



- Node color = events as % of node observations
- Line width = observations in branch as % of total observations

Node Rules



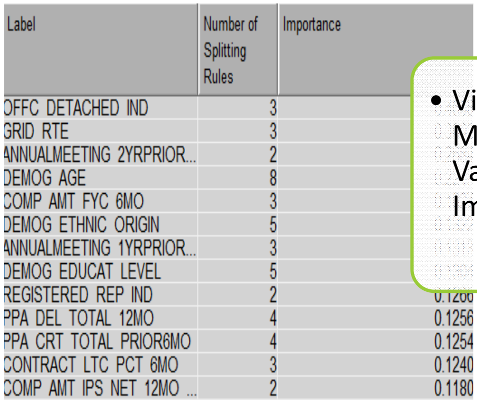
```
*------------------------------------------------------------*
 Node = 9
*------------------------------------------------------------*
if VALUE < 491496
AND DELINQ >= 5.5
AND DEBTINC < 45.1549
then
 Tree Node Identifier   = 9
 Number of Observations = 12
 Predicted: BAD = 1

*------------------------------------------------------------*
 Node = 10
*------------------------------------------------------------*
if VALUE >= 491496 or MISSING
AND REASON IS ONE OF: HOMEIMP, DEBTCON
AND DEBTINC < 45.1549
then
 Tree Node Identifier   = 10
 Number of Observations = 23
 Predicted: BAD = 1

```

- Definitions for each leaf node
- View → Model → Node Rules

Variable Importance

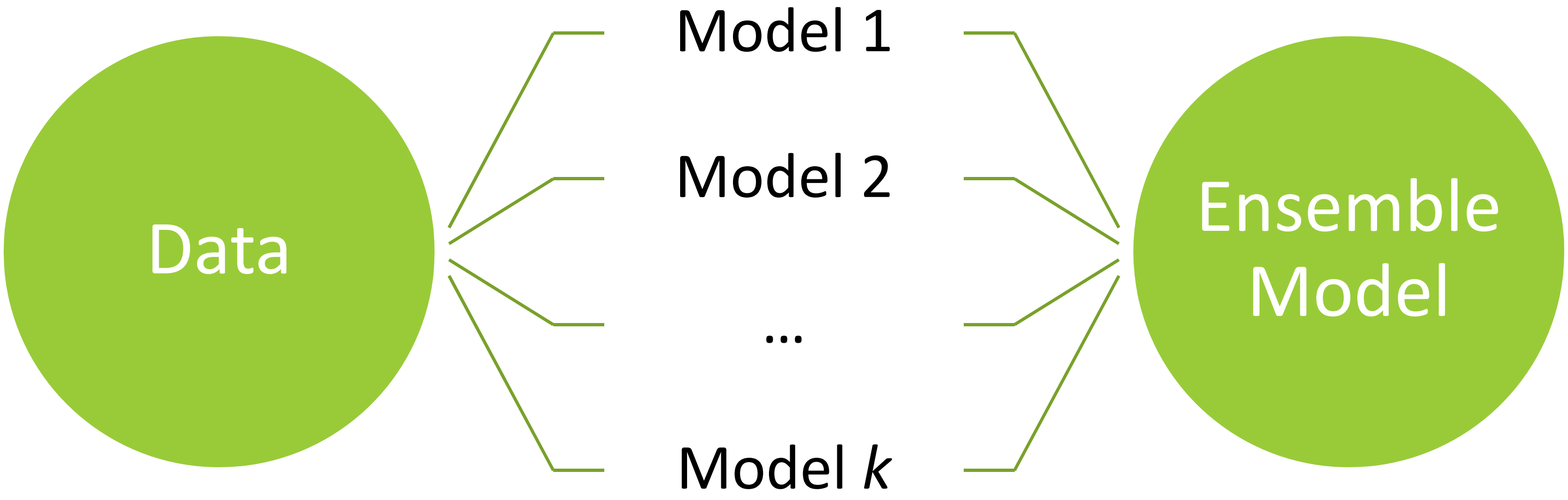


Label	Number of Splitting Rules	Importance
OFFC DETACHED IND	3	[illegible]
GRID RTE	3	[illegible]
ANNUALMEETING 2YRPRIOR...	2	[illegible]
DEMOG AGE	8	[illegible]
COMP AMT FYC 6MO	3	[illegible]
DEMOG ETHNIC ORIGIN	5	[illegible]
ANNUALMEETING 1YRPRIOR...	3	[illegible]
DEMOG EDUCAT LEVEL	5	[illegible]
REGISTERED REP IND	2	0.1266
PPA DEL TOTAL 12MO	4	0.1256
PPA CRT TOTAL PRIOR6MO	4	0.1254
CONTRACT LTC PCT 6MO	3	0.1240
COMP AMT IPS NET 12MO ...	2	0.1180

- View → Model → Variable Importance

Ensembles combine information from multiple models to produce a better, "ensembled" model

5



Often more robust and more accurate than a single model. Think "committee of experts."

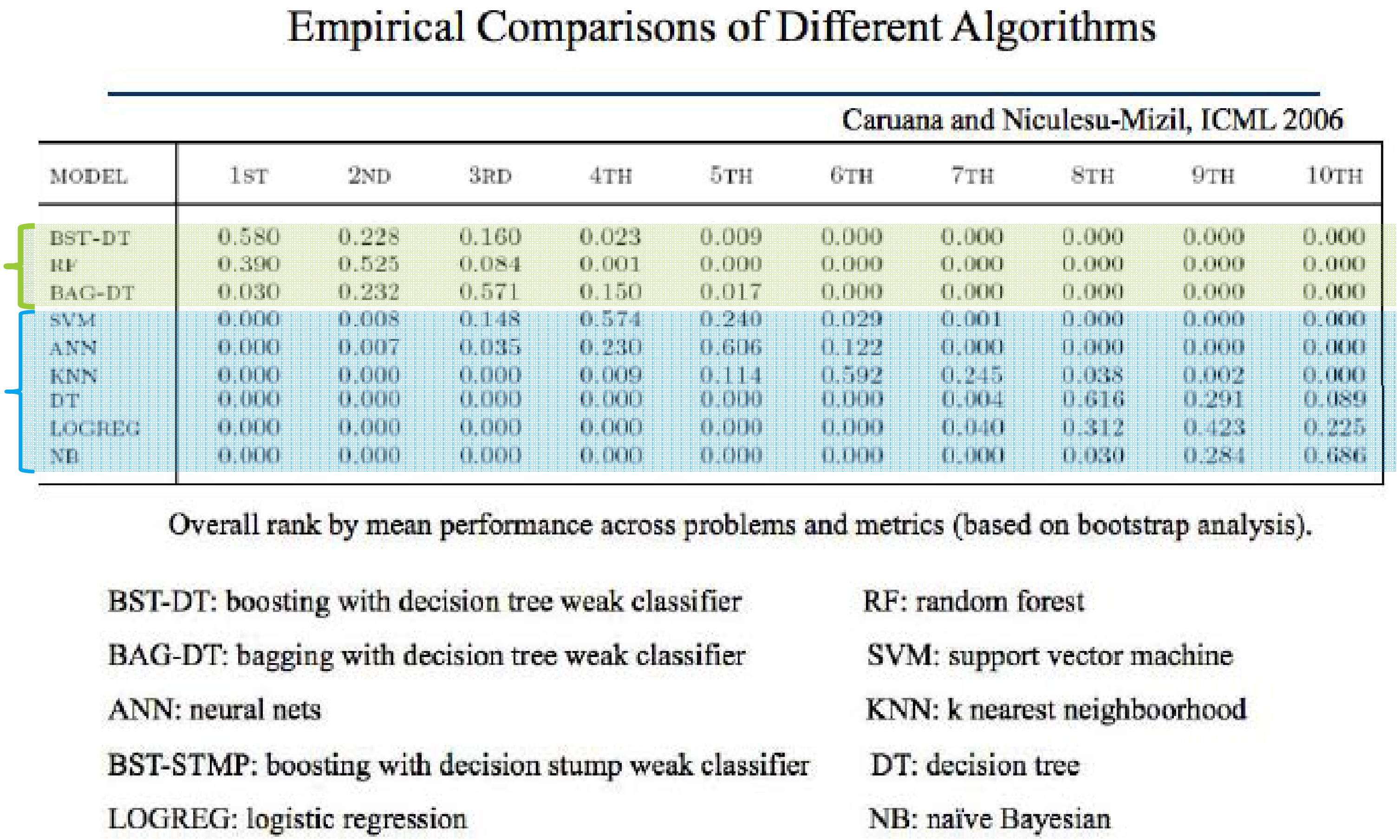
Often, but not always, ensembles are of the same algorithm, e.g. all decision tree or all neural net.

Research shows ensembles often outperform single models

5

Ensembles

Single Classifiers



Empirical Comparisons of Different Algorithms

Caruana and Niculesu-Mizil, ICML 2006

MODEL	1ST	2ND	3RD	4TH	5TH	6TH	7TH	8TH	9TH	10TH
BST-DT	0.580	0.228	0.160	0.023	0.009	0.000	0.000	0.000	0.000	0.000
RF	0.390	0.525	0.084	0.001	0.000	0.000	0.000	0.000	0.000	0.000
BAG-DT	0.030	0.232	0.571	0.150	0.017	0.000	0.000	0.000	0.000	0.000
SVM	0.000	0.008	0.148	0.574	0.240	0.029	0.001	0.000	0.000	0.000
ANN	0.000	0.007	0.035	0.230	0.606	0.122	0.000	0.000	0.000	0.000
KNN	0.000	0.000	0.000	0.009	0.114	0.592	0.245	0.038	0.002	0.000
DT	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.616	0.291	0.089
LOGREG	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.312	0.423	0.225
NB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.284	0.686

Overall rank by mean performance across problems and metrics (based on bootstrap analysis).

BST-DT: boosting with decision tree weak classifier

BAG-DT: bagging with decision tree weak classifier

ANN: neural nets

BST-STMP: boosting with decision stump weak classifier

LOGREG: logistic regression

RF: random forest

SVM: support vector machine

KNN: k nearest neighboorhood

DT: decision tree

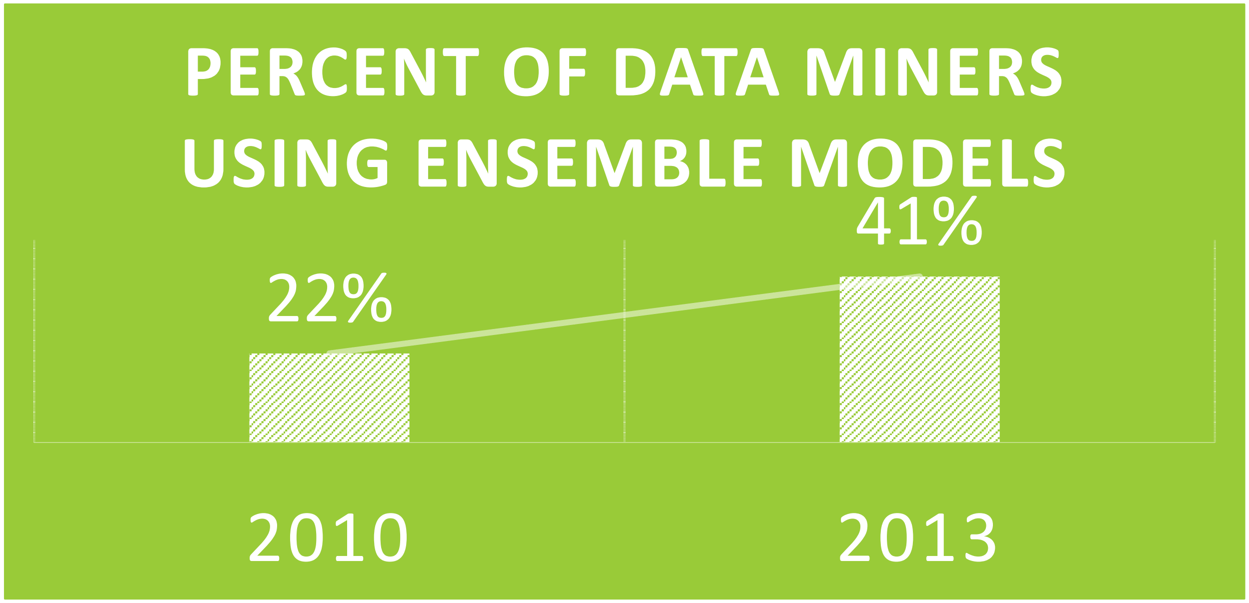
NB: naïve Bayesian

From Zhuowen Tu, "Ensemble Classification Methods: Bagging, Boosting, and Random Forests"

Ensembling is quickly becoming industry-standard

5

According to Rexer Analytics data miner survey:



Question: “What algorithms/analytic methods do you TYPICALLY use? (Select all that apply.)”

The 2013 Rexer Analytics survey included 1,200+ data miner responses from 75+ countries, and has been cited in over 60 publications.

5

Pro's and con's of ensembling

Pro's

- Can significantly reduce model error
- Easy to implement with SAS Enterprise Miner

Con's

- Model interpretability is more difficult (mitigate with Variable Importance measures)
- More computationally intensive

Rule of thumb: “good candidates” for model ensembling

5

Good candidates:
Weak learners and non-linear models

- Decision trees
- Naïve Bayes
- Neural nets
- k-Nearest Neighbor
- Support vector machine (*if* not a linear kernel; try a polynomial kernel)

Marginal candidates:
Stable, linear models

- Linear regression
- Logistic regression
- Linear discriminant analysis
- Linear support vector machines

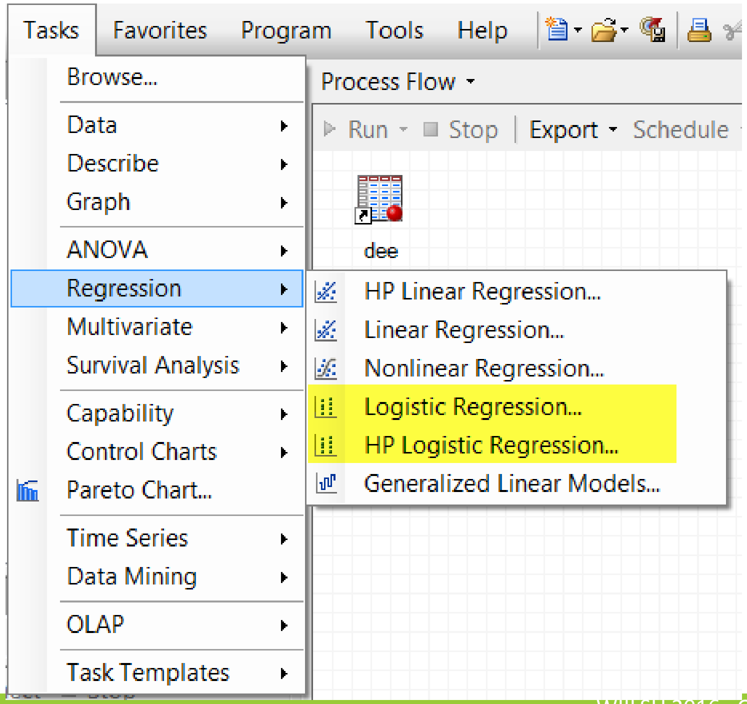
But, which algorithms work best depends on the data.

Model – How to

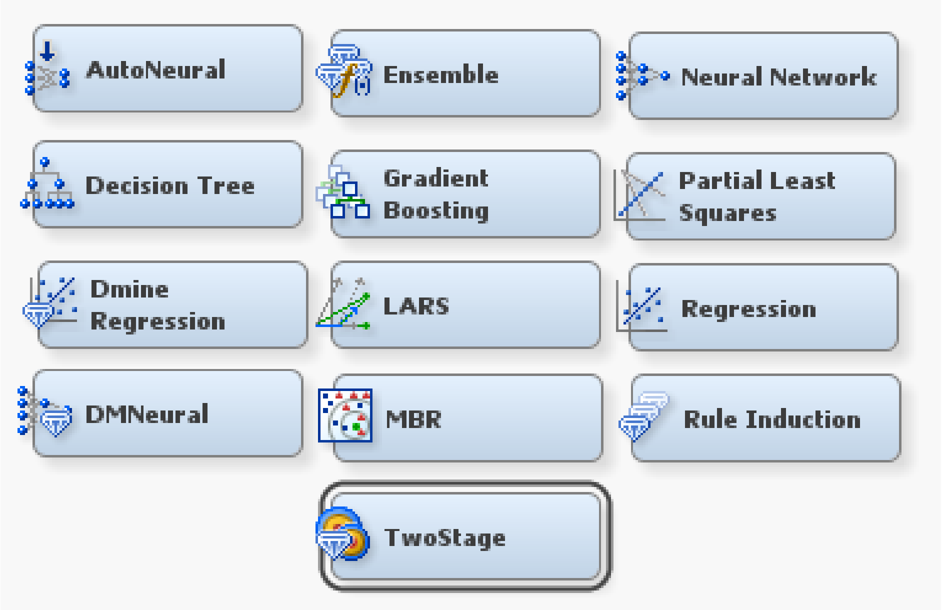
5

SAS ENTERPRISE GUIDE

- Logistic regression (forward, backward, stepwise, all-subsets)



SAS ENTERPRISE MINER



model assessment

Steps in Logistic Regression Analysis

6

Step 1: What is the probability that results could be by chance?

- Model significance (**Likelihood Ratio Test**)

Step 2: How closely do points cluster around the mean?

- Overall model strength (**Cox and Snell R^2, Nagelkerke R^2**)

Step 3: How good are the results?

- Classification Matrix (**See next Slide)**

Classification Matrix
Interpreting Logit Results

6

Predicted Result	Observed Outcome		
	0 (no)	**1 (yes)**	**Total**
0 (no)	True Negative (TN)	False Negative (FN)	TN+FN
1 (yes)	False Positive (FP)	True Positive (TP)	TP+FP
Total	TN+FP	TP+FN	TP+FP+TN+FN=N

$$Sensitivity:\ Se = \frac{TP}{TP+FN} \qquad \text{Specificity: } Sp = \frac{TN}{TN+FP}$$

$$Positive\ Predictive\ Value:\ PPV = TP/(TP+FP)$$

$$Negative\ Predictive\ Value:\ NPV = TN/(TN+FN)$$

Accuracy = (TP + TN) / N

Misclassification Rate = 1- Accuracy

Steps in Logistic Regression Analysis

6

Step 4: Are regression coefficients statistically significant?

- Significance of each model variables (**Wald Test; Odds Ratio**)

Step 5: Determining the Size of the effects

- Strength of individual contributions of **X'**s on **Y**

Step 6: Interpret standardized effects of **X'**s and rank by importance

- Standardized effects of **X'**s on **Y**
- Beta weights

Step 7: Compare Model Performance against competing Models (**AIC, BIC, or ASE**)

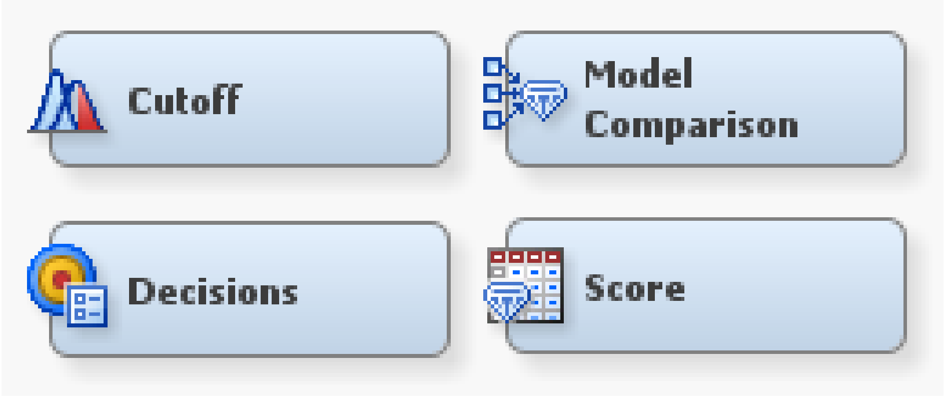
Assess – How to

6

SAS ENTERPRISE GUIDE

- PROC LOGISTIC with ROC and ROCCONTRAST statements

SAS ENTERPRISE MINER



In both platforms, look at gains charts and tables, c-statistics, (cumulative) lift, (cumulative) percent captured, misclassification rates and/or average squared error values.

Modeling Resources

Decision Trees for Analytics using SAS Enterprise Miner by de Ville & Neville

Applied Predictive Modeling by Kuhn and Johnson

Decision Trees for Predictive Modeling
(freely available online here)

Introduction to Statistical Learning
(freely available online here)

A few EDA resources

Helpful SAS EG paper with a good amount of examples including step-by-step instructions and screenshots

- http://support.sas.com/resources/papers/proceedings12/152-2012.pdf

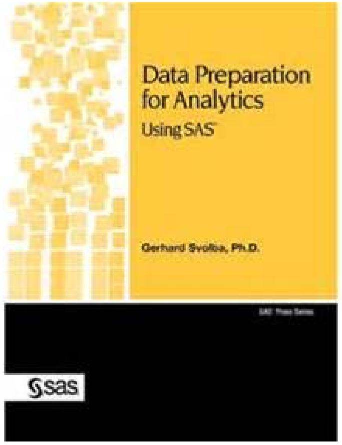
Methods for Interaction Detection in Predictive Modeling using SAS

- http://mwsug.org/proceedings/2012/SA/MWSUG-2012-SA01.pdf

Online Engineering Statistics Handbook, Section on EDA:

- http://www.itl.nist.gov/div898/handbook/eda/eda.htm

Data Preparation for Analytics – Using SAS®
by Gerhard Svolba, SAS Press Series



Contact info

Carolyn Olsen, MS GStat

Data & Analytics Consultant, Northwestern Mutual

Milwaukee, WI

Email: carolynolsen@northwesternmutual.com

Daniel Fuhrmann, Ph.D.

Data & Analytics Lead Consultant, Northwestern Mutual

Milwaukee, WI

Email: danielfuhrmann@northwesternmutual.com

The End

- Questions?
- Comments?